Task-dependent Learning of Attention

PIËRRE VAN DE LAAR, TOM HESKES, AND STAN GIELEN

RWCP (Real World Computing Partnership) Novel Functions
SNN (Foundation for Neural Networks) Laboratory
Department of Medical Physics and Biophysics, University of Nijmegen

(Received 15 January 1996; accepted 3 February 1997)

Abstract- *In this article, we propose a neural network model for selective covert visual attention. This model can learn to focus its attention on important features depending on the task to be fulfilled by gating the flow of information from the lower to the higher levels of the visual system. The model is kept as simple as possible, but it is still capable of reproducing attentional behavior observed in psychological experiments. Computer simulations demonstrate that (1) it can learn categories to reduce reaction time without a decrease in performance, (2) the model reveals a performance similar to that of humans in feature and conjunction search, and (3) its learning dynamics are comparable with those of humans.*

Keywords- Attention, Neural networks, Learning process, Visual search, Computer vision.

Acknowledgements

We would like to thank Ellen Bakker, Monique Nas, Marcel Nijman, Michel van Wanrooy, Wim Wiegerinck and two anonymous referees for their useful comments on earlier versions of this paper.

1 ATTENTION

While interacting with their environment, humans have too little time to process all available information. Therefore, only a small part of this information, which is expected to be relevant, is selected and processed in more detail. The relevance of the information depends on the task to be fulfilled. For example, paying attention to color is useful when looking out for a redbreast, but useless when trying to catch a chameleon. Psychological experiments have shown that humans can *learn* this task-dependent selection (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977). The machinery behind this adaptive selection process is called the attentional system. In this paper, we will focus on two important aspects of attention: task-dependency and adaptivity. A study of the attentional system may not only help us towards a better understanding of human

behavior but may also yield new algorithms for various applications in robotics, process control and computer vision.

Inspired by Johnston and Dark (1986) we define attention as the differential processing of simultaneous sources of information. These sources can be internal (memory and knowledge) as well as external (environmental objects and events). There are many ways in which attention can be divided. Before giving the exact definitions of the various aspects of attention relevant for this article, we will first illustrate them using the well-known cocktail party example, see e.g. von der Malsburg and Schneider (1986).

During a cocktail party many conversations take place at the same time. Despite this cacophony, humans are able to follow a single conversation (*selective attention*) and with some effort they can even follow two conversations simultaneously (*divided attention*). The mentioning of one's name or a loud bang can draw one's attention away from the current conversation towards the source of this sound (*exogenous attention*). Another change of conversation may be caused by our own intentions (*endogenous attention*). For example, we can walk away from an uninteresting speaker (*overt attention*), or we can switch between conversations without any movement (*covert attention*).

Attention is used to reduce the amount of information to be handled by the higher cognitive levels in the brain while maintaining a high performance. Selective attention processes only one source of information, whereas divided attention simultaneously processes more than one source of information (not necessarily all) up to the highest cognitive levels.

An object may attract attention based on two different grounds: exogenous or endogenous. Objects that greatly differ from their environment in, for example, color, closure, size, or three-dimensional orientation (Enns, 1990; Treisman, 1986), and objects that are of special relevance such as one's name, attract exogenous or bottom-up attention [1]. In endogenous or top-down attention the higher cognitive levels in the brain influence the attentional system to bias the selection in favor of a particular (combination of) feature(s). This selection can, of course, only be biased on the basis of information available at the time this decision is made. In other words, only information about the object which is preattentively extracted can be used to change the preferences of the attentional system.

The last relevant division of attention is that into overt and covert attention. Overt attention refers to observable movement, e.g. head and eye movement, to change the focus of attention. A well-known example is the depiction by Yarbus (1967) of the eye movements made by a subject scanning a visual scene. Covert attention, on the other hand, changes the gaze without any motor action. Covert attention is used, for example, to read the characters on an acuity chart (Anstis, 1974) while fixating the center.

In the literature on selective visual attention, several models have been proposed. For example, feature-integration theory (Treisman and Gelade, 1980), SLAM (Phaf et al., 1990),

[1] Exogenous attention is subject to learning. See, for example, Schneider and Shiffrin (1977) for the learning of automatic responses to visual stimuli. The automatic response to one's name shows that exogenous attention can also be learned for auditory stimuli, since without learning, attention should be exogenously captured by either all names or none at all.

VISIT (Ahmad, 1992), the neocognitron (Fukushima and Imagawa, 1993), dynamic routing circuits (Olshausen et al., 1993), SCAN (Postma, 1994), SERR (Humphreys and Müller, 1993) and guided search (Wolfe, 1994). Whereas humans learn and use attention in multiple tasks simultaneously, these models are either hard-wired without learning properties or perform just one specific task. This study, therefore, will focus on the *task-dependent learning* of selective covert visual attention, both exogenous and endogenous. A neural network which receives task-dependent input will be used to gate the flow of information in the preattentive stage. Thanks to this gating, the model can learn to direct its focus of attention towards locations containing relevant information for the task to be performed.

In Section 2 we will describe our model. In Section 3 we will present simulations of visual search tasks using our model and we will compare its performance with psychological data. In the general discussion (Section 4) we will compare the characteristic features and the performance of our model with that of previously published models. Moreover, we will also describe some other psychological data that can also be explained and we will indicate some possible extensions.

2 THE MODEL

In this section we describe our neural network model for selective covert visual attention. The attentional mechanism of our model only selects between external sources. Internal sources, e.g. thoughts or memory, are thus ignored. We will keep our model as simple as possible, while still being able to reproduce attentional behavior observed in psychological experiments.

The human attentional system can be divided into two sequential stages: a preattentive stage and a limited-capacity stage (Treisman, 1986; Wolfe, 1994). In the preattentive stage the information at each spatial position is processed in parallel. This stage determines the relevance of the stimulus features at the different locations of the visual field for the current task. Based on this relevance assessment, the limited-capacity stage is assumed to perform a serial self-terminating search.

2.1 Preattentive stage

The preattentive stage of our model is depicted in Figure 1. This stage consists of a sensory input comparable to the retina's input, feature[2] maps, an attentional network, and a priority map. The model's "retina" views the static images. This retina has a homogeneous distribution of receptors. These receptors pass their information on to the feature maps in accordance with

[2] Like Treisman and Sato (1990) we define "dimension" as a complete set of mutually exclusive values, at least one of which must characterize any stimulus to which the dimension applies and "feature" as a value on such a dimension (for example, "green" on the color dimension; "horizontal" on the orientation dimension).

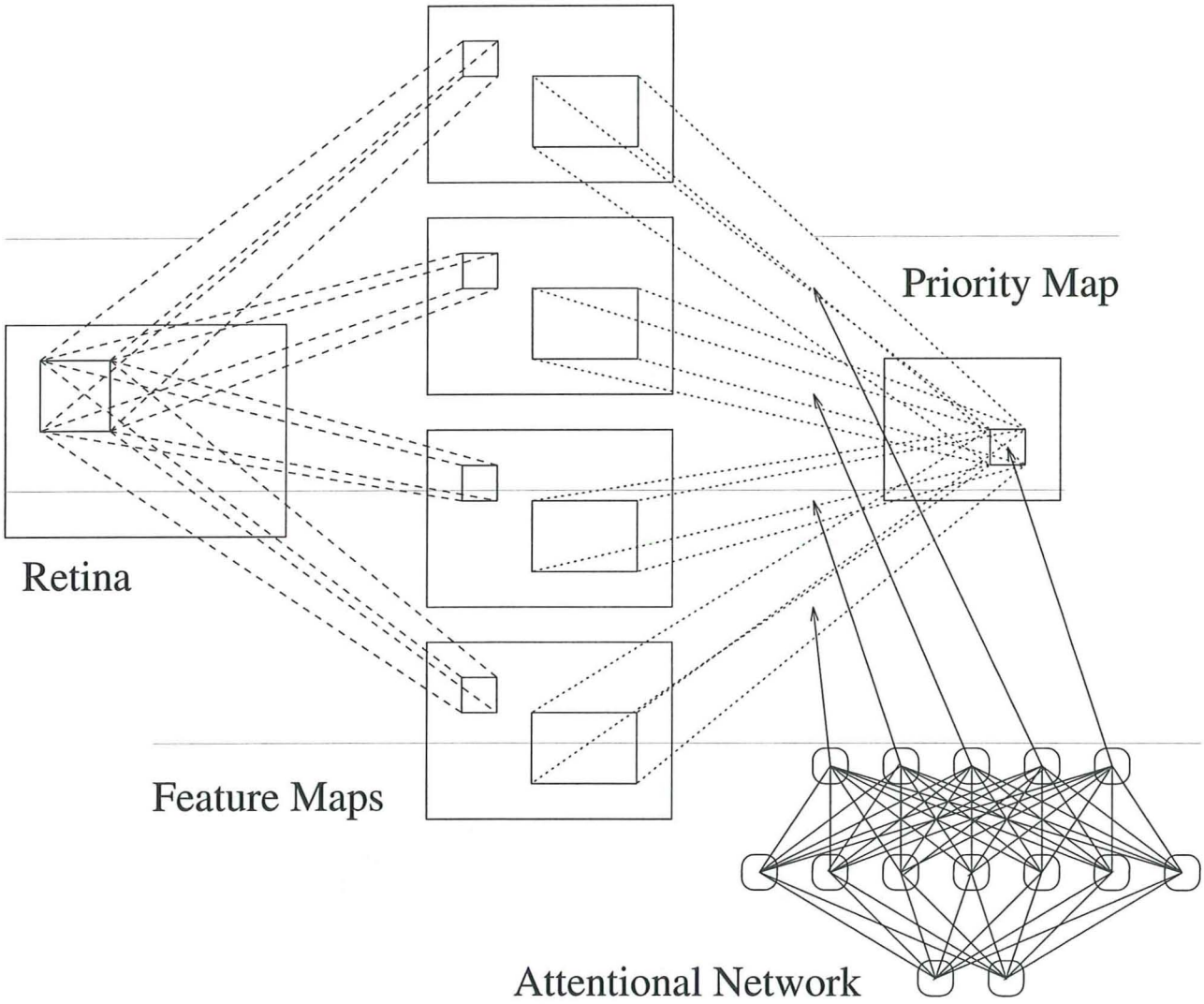


FIGURE 1. The architecture of the artificial visual system. From left to right: the retina, the feature maps, the attentional network, and the priority map. The information going from the feature maps to the priority map is influenced by the attentional network, which receives a task-dependent input.

the following formula

$$i^{\theta}_{\kappa\lambda} = \sum_{\mu=0}^{D_x-1} \sum_{\nu=0}^{D_y-1} d^{\theta}_{\mu\nu} r_{\kappa+\mu,\lambda+\nu} + \xi \,, \tag{1}$$

with $r_{\kappa\lambda}$ the activity of the retina at location (κ, λ), ξ Gaussian noise, D_x (D_y) the number of receptors in the receptive field of the neurons in the feature map in the horizontal (vertical) direction, $i^{\theta}_{\kappa\lambda}$ the input of the neuron (κ, λ) of this feature map, and $d^{\theta}_{\mu\nu}$ are shared weights of feature map θ. Weight sharing (Rumelhart et al., 1986; LeCun et al., 1990) constraints all neurons in a map to compute exactly the same function over their receptive fields. Consequently, the response of a neuron does not depend on its position in the map, but only on the objects in its receptive field. Because all neurons in a map have the same optimal stimulus, the map can be described by this optimal stimulus, for example a red blob, a green blob, a horizontally oriented bar or a vertically oriented bar. Thanks to weight sharing, translation of a stimulus only results in translation of activity within the maps. Neurons in a feature map are topologically ordered,

meaning that neighboring neurons have neighboring receptive fields. The activity of the neurons in a feature map is normalized and depends on the total input of this feature map:

$$f^{\theta}_{\kappa\lambda} = \frac{i^{\theta}_{\kappa\lambda}}{\text{MAX}\left(1, \sum_{\mu\nu} \left|i^{\theta}_{\mu\nu}\right|\right)} , \tag{2}$$

where $f^{\theta}_{\kappa\lambda}$ represents the activity of the neuron (κ, λ) of the feature map θ. The MAX function prevents noise amplification in the case of absence of input to a feature map. When only one neuron in a feature map receives input, normalization has "no effect" and the stimulated neuron will respond maximally. When several neurons in a feature map receive input, normalization will reduce the responses of the stimulated neurons such that the total activity of the feature map will be constant. Due to the normalization, the total activity of the feature map is independent of the number of stimulated neurons. Related to that, the response of a neuron to a feature does depend on the presence of the same feature in other parts in the visual field.

The information in the various feature maps is passed on to the priority map. The activity in the priority map represents the relevance of the different parts of the visual field. The flow of information from the feature maps to the priority map is modulated by the output of the attentional network. The activity of neuron (κ, λ) in the priority map $p_{\kappa\lambda}$ is given by

$$p_{\kappa\lambda} = \tanh\left(\sum_{\theta} \sum_{\mu=0}^{A_x-1} \sum_{\nu=0}^{A_y-1} a^{\theta}_{\mu\nu} f^{\theta}_{\kappa+\mu,\lambda+\nu} + \xi\right) , \tag{3}$$

with A_x (A_y) the number of receptors in the receptive field of the neurons in the priority map in the horizontal (vertical) direction, and $a^{\theta}_{\mu\nu}$ the attentional shared weights which are the output of the attentional network.

The attentional network, which receives task-dependent input, is responsible for the top-down control of attention by affecting the flow of information from the lower levels to the higher levels of the visual system. The output of the attentional network reads

$$a_i = a_i(\boldsymbol{m}, \boldsymbol{w}) , \tag{4}$$

where $\boldsymbol{m}$ denotes the memory input (also called task input); a_i is the i-th output of the attentional network; i is short for (θ, μ, ν); and $\boldsymbol{w}$ represents the weights in the attentional network.

The input of our model is a visual image at the retina and a task to be performed at the input of the attentional system. The model is trained by gradient descent to minimize the following energy function

$$E(d^{\theta}_{\mu\nu}, \boldsymbol{w}) = \sum_{q} \sum_{\kappa\lambda} \left(p_{\kappa\lambda}(\boldsymbol{r}^q, \boldsymbol{m}^q, d^{\theta}_{\mu\nu}, \boldsymbol{w}) - t^q_{\kappa\lambda}\right)^2 \tag{5}$$

where E is the energy function, q labels the patterns and $t_{\kappa\lambda}$ is the desired output for neuron (κ, λ) in the priority map. The desired output at the target's location is maximal activity $(t_{\kappa\lambda} = 1)$ and minimal activity everywhere else $(t_{\kappa\lambda} = -1)$. For example, assume that the task is to find a red horizontal bar. A possible visual image $\boldsymbol{r}$ contains a number of green horizontal and red vertical bars and one red horizontal bar, the target of this search task. The task is represented in the input of the attentional system $\boldsymbol{m}$. The desired activity $\boldsymbol{t}$ in the priority

map is the maximal activity at the location of the red horizontal bar and the minimal activity everywhere else. During training, the difference between the desired activity and the actual activity for inputs $\boldsymbol{r}$ and $\boldsymbol{m}$ is used as the error to be backpropagated.

In all simulations, described in this article, the weights $d^{\theta}_{\mu\nu}$ of the feature maps are fixed and only the weights of the attentional network are learned. The weights $d^{\theta}_{\mu\nu}$ are set to yield feature maps (e.g. luminance, orientation, and size) which are loosely comparable to those used by humans. For the attentional network a feedforward perceptron is used which updates the weight vector $\boldsymbol{w}$ by the learning rule,

$$\Delta \boldsymbol{w} = -\eta \frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} \tag{6}$$

with η the learning parameter. Note that the output of the attentional network $\boldsymbol{a}$, which serves as the weights for gating the output of the feature maps, depends on the weights $\boldsymbol{w}$ and therefore, is learned by training $\boldsymbol{w}$. The output of the attentional network $\boldsymbol{a}$ was never imposed during learning. The number of output neurons of the attentional network is equal to the number of feature maps plus one for a threshold for the neurons in the priority map. Furthermore, the noise is randomly drawn from a Gaussian distribution with a standard deviation of $\sigma = 0.1$. This noise not only makes our model more realistic from a neurobiological point of view, but also gives our model a nondeterministic behavior necessary to explain data observed in psychological experiments.

2.2 Limited-capacity stage

In the limited-capacity stage, first, the locus of attention is determined by a saliency map. All neurons in this map interact through local excitation and global inhibition with each other such that only one blob of activity is possible (Amari, 1977; Glasius et al., 1994). When the saliency map has converged to a stable state, the blob of activity will be at the location with the largest external input, similarly to a "winner-take-all" mechanism. This location is the locus of attention. Then only the information inside the focus of attention is further processed (Leow and Miikkulainen, 1991; Olshausen et al., 1993; Postma, 1994) to see whether it indeed contains a target. If so, the search ends, otherwise, the current location is inhibited (Klein, 1988; Niebur and Koch, 1996) and a new blob of activity will appear in the map at the location corresponding to the new largest external input. This new blob determines the new locus of attention. The model repeats these steps either until a target is found or until all positions have been visited.

In order to be able to compare the model's performance with human performance we have to specify the values of the initial response time, i.e., the time needed just to respond, and the duration of a shift of attention, i.e., the time needed to move attention from the current location to the new position. Based on values extracted from psychological experiments (Treisman and Sato, 1990; Wolfe et al., 1989), the initial response time has been set to 300 ms and the duration of a shift of attention has been set to 60 ms. With these definitions, the model's reaction time is defined as

$$RT = 300 + 60 S_a \tag{7}$$

where S_a is the number of shifts of attention made by the model while searching for the target and RT is the reaction time of the model (in ms).

3 VISUAL SEARCH TASKS

In a typical psychological experiment of visual search the subject is instructed to look at a screen. On this screen one frame or a sequence of frames is presented. A frame contains randomly [3] positioned items, the so-called frame items. The number of items in a frame is called the frame size. Before the presentation of a frame (or sequence of frames), a memory set is shown to the subject. The memory set contains all possible targets of the search. Distractors are all items that appear in a frame without being part of the memory set. The task of the subject is to report whether an object from the memory set did or did not occur in (at least one of) the frame(s). Reaction time and accuracy are measured as a function of the frame size.

For our simulations, we have created frames similar to those used in the psychological experiments. Each frame contains one randomly positioned target. Each remaining position in the frame is filled by a distractor, which is randomly chosen out of all possible distractors, given the task and the target. These frames define the visual input of the model. The memory input defines the goal of the search task, and therefore depends on the task and the target.

Although humans may know discriminating features between the target and the distractors in a search task, they may not be able to use this information in the preattentive stage of the attentional visual system. For example, despite knowing that an H has one more horizontal bar than an H, a subject is unable to detect the H between H's in parallel [4] (Quinlan and Humphreys, 1987). To make the simulations as realistic as possible, the memory input of the attentional network contained only information which humans can use preattentively to bias their selection of information. Of course, for various applications in robotics, process control and computer vision this limitation does not have to be imposed and the model would outperform humans.

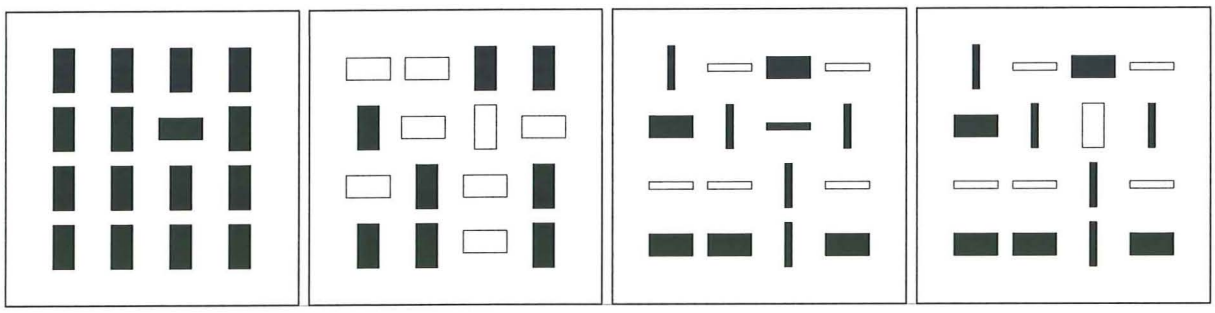
First, we will test whether our model has a performance similar to that of humans in feature and conjunction search after training. Second, the model will be tested on a category learning paradigm. Finally, its learning dynamics will be compared with human data.

3.1 Feature and conjunction search

A visual search paradigm that is often used (Quinlan and Humphreys, 1987; Treisman and Sato, 1990; Wolfe et al., 1989) is feature and conjunction search. In a *feature search* task the target differs in one and the same feature from all distractors. See Figure 2(a), for example. In a

[3] The positions of the items are not completely random. The items do not overlap and are normally arranged in regular rectangular matrices or on the perimeter of an imaginary circle.

[4] A target is defined to be detected in parallel when the slope of the reaction time as a function of the frame size is below 10 ms/item (Treisman and Sato, 1990). Otherwise, the target is said to be detected by a serial search.



(a) (b) (c) (d)

FIGURE 2. An example of feature search (a), standard conjunction search (b), triple conjunction search with the target differing from all distractors in one (c) and two (d) features.

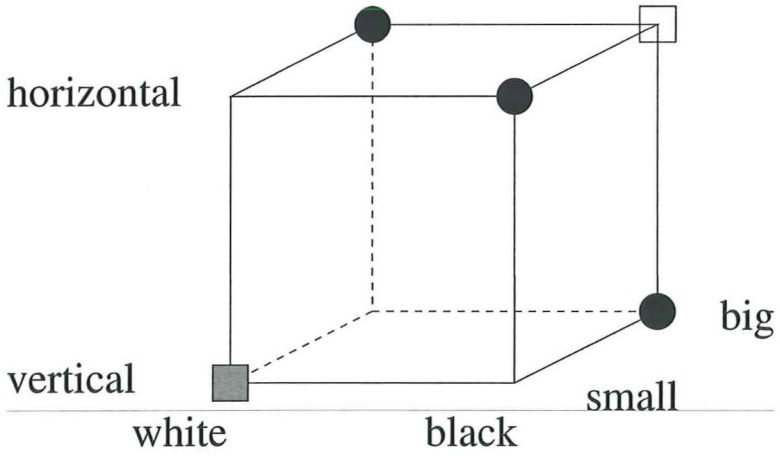


FIGURE 3. The feature space of a search task. The three black dots symbolize a possible set of distractors in a triple conjunction search and the white and gray squares symbolize the corresponding targets, which differ from all distractors in one and two features, respectively.

conjunction search task the item to be found differs from each distractor group, but not from all distractors in the same feature. In the *standard* conjunction search, as depicted in Figure 2(b), there are two distractor groups, each possessing two features, and the target shares one feature with each distractor group. A *triple* conjunction search task has two special cases (see also Figure 3). The target may differ from all distractor groups in one feature, as in Figure 2(c), or in two features, as in Figure 2(d).

3.1.1 Human performance

All studies on feature and conjunction search (Quinlan and Humphreys, 1987; Treisman and Sato, 1990; Wolfe et al., 1989) report that in feature search the target is detected in parallel. However, these studies claim different human performances in conjunction search. The experiments done by Wolfe et al. (1989) and Treisman and Sato (1990) suggest that the standard conjunction target is detected in parallel. Quinlan and Humphreys (1987), on the other hand, state that subjects perform a serial search to detect the target. Our model will be tested on

Slope of the reaction time as a function of the frame size (ms/item)	Human	Model
Feature search	2.2	1.7
Standard conjunction search	30.6	29.4
Triple conjunction search; target differs in one feature	37.3	36.7
Triple conjunction search; target differs in two features	11.6	16.0

TABLE 1
Slope of reaction time as a function of the frame size determined by Quinlan and Humphreys (1987) (the slopes of feature and standard conjunction search are calculated by averaging over the experiments one, two, and three) and produced by the model.

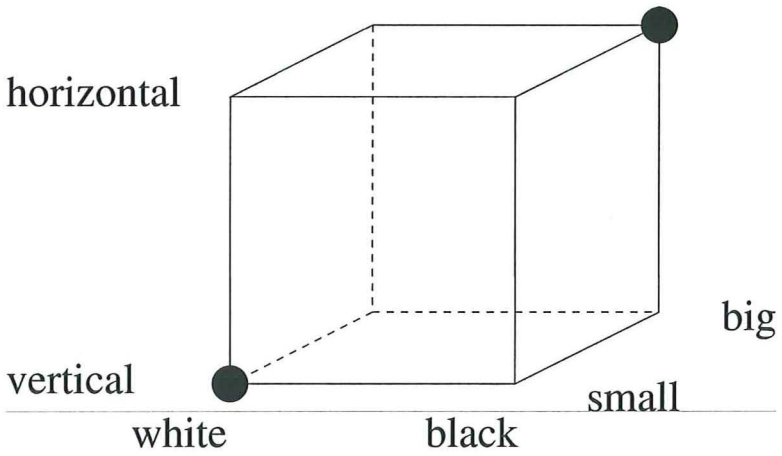


FIGURE 4. The feature space of a conjunction search task. The black dots symbolize a (possible) pair of maximally contrasting stimuli. When this pair of stimuli is used as distractors, six conjunction targets are possible as indicated by the remaining corners of the cube.

both paradigms.

The first experiment we describe has been designed by Quinlan and Humphreys (1987). They report that in feature search the target is detected in parallel across the visual field. Furthermore, they show that the reaction time in both standard conjunction search and triple conjunction search is a linear function of the frame size. However, when the triple conjunction target differs from all distractors in one feature (two features), the slope is steeper (flatter) than in the case of standard conjunction search. See also Table 1.

The second simulation is based on an experiment conducted by Treisman and Sato (1990), in which the impact of top-down information can be manipulated. In this experiment for each subject a pair of maximally contrasting distractors is randomly chosen. For each pair of distractors, there are six possible conjunction targets (see also Figure 4). Subjects are asked to search either for one specific target or for one of these six targets. The memory set thus contains one item or six items, respectively. Treisman and Sato call these two paradigms the *known* and *unknown* case to reflect the subject's knowledge about the specific target.

To illustrate how the top-down information is manipulated consider, for example, a subject who

has to find a conjunction target between distractors. The distractor set consists of a small white vertically oriented bar and a big black horizontally oriented bar. As can be seen in Figure 4, there are six possible targets:

- the small black horizontally oriented bar;
- the small black vertically oriented bar;
- the small white horizontally oriented bar;
- the big white vertically oriented bar;
- the big white horizontally oriented bar; and
- the big black vertically oriented bar.

Because humans do not have maps in the preattentive stage for combinations of features of the color, orientation, and size dimensions (Quinlan and Humphreys, 1987; Treisman and Sato, 1990; Wolfe et al., 1989), the target and distractors must be separated purely based on features, not on conjunctions of features. Therefore, humans use top-down information to select features that separates the target from (some or all) distractors, while ignoring features that are shared by both the target and all distractors. As can be easily seen from the above list, in triple conjunction search for one specific target the target has activity in all three corresponding feature maps. However, the distractors have activity in only one or two of these feature maps. This difference can be used to direct the focus of attention towards the target and thus leads to a shorter reaction time. On the other hand, in a search for one out of six conjunction targets, no such combination of feature maps can be found since the information contained in each feature map is as beneficial to three possible targets as it is harmful to the other three possible targets. Therefore, whereas in the known case top-down information can be used to reduce the reaction time, it is useless in the unknown case.

In the experiments of Treisman and Sato (1990), the slope of the reaction time as a function of the frame size is steep in the unknown case. However, the reaction time is (almost) independent of frame size when the specific target is known.

3.1.2 Simulations

The frames used in the simulations are of the type shown in Figure 2. The eight possible frame items are either black or white, either horizontally oriented or vertically oriented, and either big or small (see also Figure 3). Depending on the search task, the stimulus consists of a target and corresponding distractors selected from these eight items. The model has six feature maps, representing the possible features of the frame items.

Our first simulation concerns the detection of a target in the paradigm of Quinlan and Humphreys (1987). During this simulation, the task input is held constant because humans do not have feature maps for H's and H's. Therefore, no top-down information can be used in conjunction search. In feature search, we also assume that attention is purely exogenously driven[5]. We train the model simultaneously on feature, standard conjunction, and triple conjunction

[5] When subjects use top-down information in feature search, the assumption of constant task input is invalid. But the target is still detected in parallel across the frame, although the subjects' reaction time will be faster (Treisman, 1986).

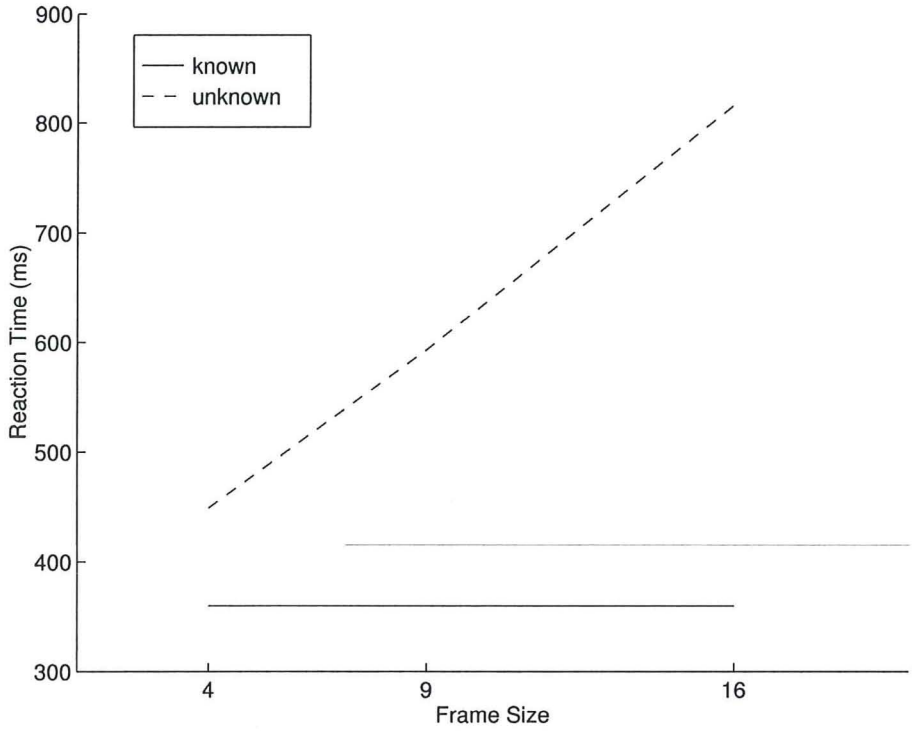


FIGURE 5. Search times for known and unknown conjunction targets.

search with the target being different from the distractors in one and two features, respectively. We prefer this training scheme, since other training schemes yield performances different from humans. For example, when the model has been trained on only triple conjunction search with the target different in one feature, its performance on this search task is better than humans, but it is worse on all other search tasks. This difference in performances is due to the different bottom-up activation of the targets in the feature and conjunction searches (Cave and Wolfe, 1990). After training, we determine the slopes of the reaction time as a function of frame size. These slopes are displayed on the right side of Table 1.

The second simulation resembles an experiment done by Treisman and Sato (1990). The task input consists of six nodes; one node for each feature. When a feature is present in the target or in one of the possible targets, the value of the corresponding task input node is equal to one and otherwise minus one. We randomly select two distractors (see Figure 4 for an example) and we use these distractors to train the model simultaneously in both the known and the unknown case. Figure 5 shows the performance after training.

3.1.3 Discussion

After training, our model has a performance similar to that of humans. Like the subjects in the experiments of Quinlan and Humphreys (1987) the target is detected in parallel across the visual field in feature search. In both standard and triple conjunction search there is a linear relation between reaction time and frame size. The slope when the triple conjunction target differs from all distractors in one feature (two features) is steeper (flatter) than in standard conjunction search. Like Treisman and Sato (1990) we find that task-dependencies do indeed influence the performance. In the case of a known target the search is parallel, whereas in the

Human		
Target	Inter	Intra
Non-common	370	410
Common	380	400

Model		
Target	Inter	Intra
Non-common	361	433
Common	363	387

TABLE 2
Reaction time (ms) to detect a known target between three distractors. The left table contains the results of the experiment done by Jonides and Gleitman (1972) The values determined with our model, after learning, are in the right table.

case of an unknown target the search appears to be serial. This simulation shows that our model is able to learn to focus its attention on task relevant features. This ability reduces the reaction time while maintaining its performance.

The difference in performance in conjunction search between the various studies (Quinlan and Humphreys, 1987; Treisman and Sato, 1990; Wolfe et al., 1989) can be explained in at least three ways. This difference may be due to different levels of noise across the subjects (Cave and Wolfe, 1990) or to a difference in stimulus salience between the tasks (Wolfe et al., 1989). Yet, as our simulations show, the usefulness of top-down information seems to be the most obvious explanatory factor for the discrepancy between the different studies. Although a subject might know a discriminating feature or combination of features which distinguish the target from the distractor(s), information about the (combination of) feature(s) need not be preattentively available to make the division between target and distractors in parallel, and consequently, a serial search is performed. For example, where *O* and *X* can be preattentively distinguished on the basis of closure and oriented bars, H and H cannot be separated because both are made up of horizontally and vertically oriented bars.

3.2 Categories

In some endogenous visual search tasks the target and the distractors are elements of categories, e.g. letters or digits. Using these categories, we can divide the visual search into *intercategorical* and *intracategorical* search. When performing an intercategorical search the subject has to find an item from one category between distractors, which are all elements of another category. In intracategorical search the subject has to look for a target which is a member of the same category as all distractors. For example, seeking for an *A* between {*1*,*4*,*9*} is an intercategorical search, but between {*B*,*D*,*H*} an intracategorical search.

3.2.1 Human performance

In the paradigm of Jonides and Gleitman (1972), the subjects have to search for a verbally specified object between distractors. These distractors are either all letters or all digits. One of the objects to appear in a frame is the ambiguous symbol *O*. This oval shape can be interpreted either as the letter *o* or as the digit **0**.

In the Jonides and Gleitman (1972) experiment the reaction time is measured when looking for

- a member of category X, not equal to O, between members of category Y (intercategorical search for non-common element);
- a member of category X, not equal to O, between members of category X (intracategorical search for non-common element);
- O specified as a member of category X between members of category Y (intercategorical search for common element);
- O specified as a member of category X between members of category X (intracategorical search for common element);

where X stands for letters and Y for digits, or vice versa. The results of the Jonides and Gleitman study are displayed on the left side of Table 2.

The experimental results indicate that humans need more time to detect a target among members of its own category than among members of another category. An even more striking result is that subjects respond faster (more slowly) when they search between digits (letters) for the letter o than when they search for the digit $\emptyset$, although the same symbol (O) is the target of both search tasks. This result clearly shows that knowledge about the category affects the attentional system.

3.2.2 Simulations

To test our model on the paradigm of Jonides and Gleitman (1972) we define a similar, yet simpler problem. We assume that there are nine possible features. We define two categories, A and B, whose objects have three features out of nine. Denoting the features as x_α with $\alpha \in [1, 2, .., 9]$ the two categories are

$$A = \{\, (y_1, y_2, y_3) \,\} \qquad \text{with} \quad y_\beta \in \{x_\beta, x_{\beta+3}\}\ ,$$

and

$$B = \{\, (z_1, z_2, z_3) \,\} \qquad \text{with} \quad z_\beta \in \{x_\beta, x_{\beta+6}\}\ .$$

These two categories have eight elements each. One object $[(1, 2, 3)]$, having the features one, two, and three, is a member of both categories. This element corresponds to the symbol O in the experimental paradigm. In the simulations of this paradigm, there are nine feature maps. The model is first trained to learn the categories. During learning the number of intercategorical searches was seven times higher than the number of intracategorical searches. The target of the search is partly defined by the two memory inputs. One memory input reflects the category of the target, the other reflects the O-ness of the target. Searching for the letter o is thus different from searching for the digit zero, although the target of both searches is the same ambiguous symbol O. The attentional network has only one hidden unit. This limitation reflects the difficulty of the task and prevents the network from learning four categories: letters, digits, O's and not-O's. The reaction times of the model after training are shown in Table 2.

3.2.3 Discussion

The reaction times of the model are similar to those measured by Jonides and Gleitman (1972). A target is detected faster among members of another category than among members of its own category. Furthermore, the reaction time is shorter in a search for a common element specified as a member of category X between members of category Y than between members of category X. Our model is therefore able to learn categories and to use these categories to reduce the reaction time in the various visual search tasks.

3.3 Learning and unlearning

The subjects in the experiments of Schneider and Shiffrin (1977) have to search a sequence of frames for memory set items. The items in the frames are selected from two disjunct character sets. In consistent mapping conditions one of these sets is always the target set and the other is always the distractor set. In varied mapping conditions the target and distractors are never of the same set, but the target is randomly chosen amongst the sets.

In the learning and unlearning experiment the subjects have to search a sequence of frames, all of size two, for four memory set items. The items in the frames are selected from two disjunct character sets $\{B,C,D,F,G,H,J,K,L\}$ and $\{Q,R,S,T,V,W,X,Y,Z\}$. The subjects' performance is first determined in 2100 sequences of frames using consistent mapping. After these 2100 sequences the target and distractor sets are exchanged. What used to be a target becomes a distractor and vice versa. The performance of the subjects is again determined for an additional 2400 sequences of frames.

3.3.1 Human performance

Human performance after training under consistent and varied mapping conditions differs substantially. After training under consistent mapping conditions, the subjects are able to detect the target in parallel, but after training under varied mapping conditions they have not developed this ability.

The performance of the human subjects in the learning and unlearning experiment can best be described as follows. When targets and distractors are selected from two disjunct fixed sets, the performance of the subjects improves over time. However, when the target and distractor sets are reversed, the performance of these "trained" subjects drops far below the initial untrained level. Of course, continual training again leads to a better performance, but it takes a longer period of time to achieve the same accuracy as before reversal.

3.3.2 Simulations

During the simulations our model "looks" at one frame with two objects, rather than at a sequence of frames. The presentation time is short such that the model can only look at one position of the frame. After each trial the model is informed about the target location. We use the same target and distractor sets as Schneider and Shiffrin. The feature maps encode closure,

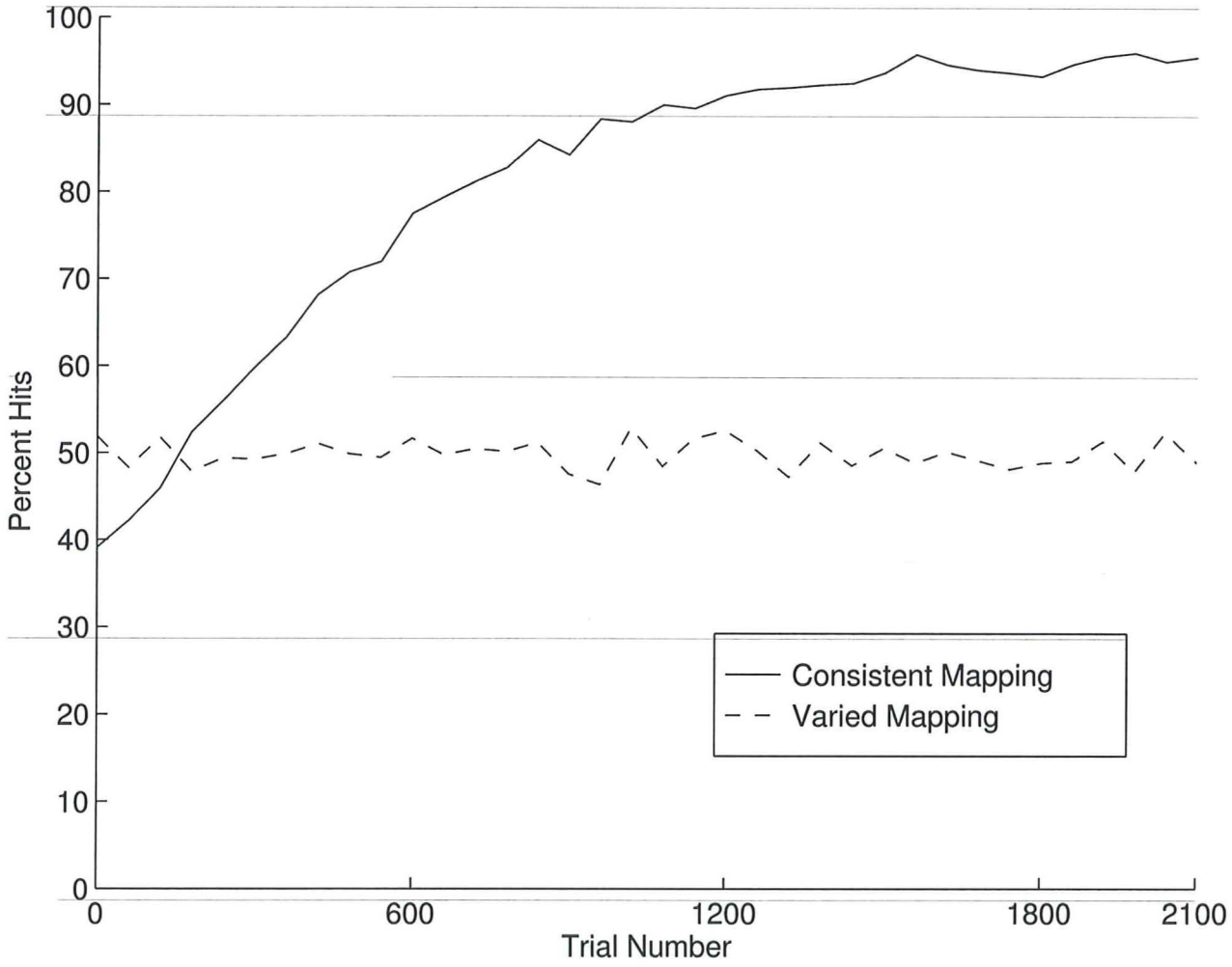


FIGURE 6. The learning dynamics in the case of consistent and varied mapping.

bars of particular orientation in particular parts of the receptive field and symmetries. The task input is held constant during all trials for two reasons. First, the ability to detect the targets in parallel by the subjects appears to be task-independent. For example, after being a subject in a consistent mapping experiment, reading is hindered because the memory-set items "jump out" from the page [p. 153 of (Shiffrin and Schneider, 1977)]. Second, when we do train our model with task input specifying the possible targets of the search, the model detects the target in parallel, even under varied mapping conditions (unlike humans).

In the first simulation we determine the model's performance in 2100 learning trials under consistent and varied mapping conditions. See Figure 6 for the results. In the second simulation we measure the performance in the learning and unlearning experiment. The learning dynamics are depicted in Figure 7.

3.3.3 Discussion

On the basis of their results, Schneider and Shiffrin (1977) conclude that humans use two processes, controlled search and automatic detection, which operate in parallel in visual search. They define controlled search as a serial process which is easily established, altered, and even

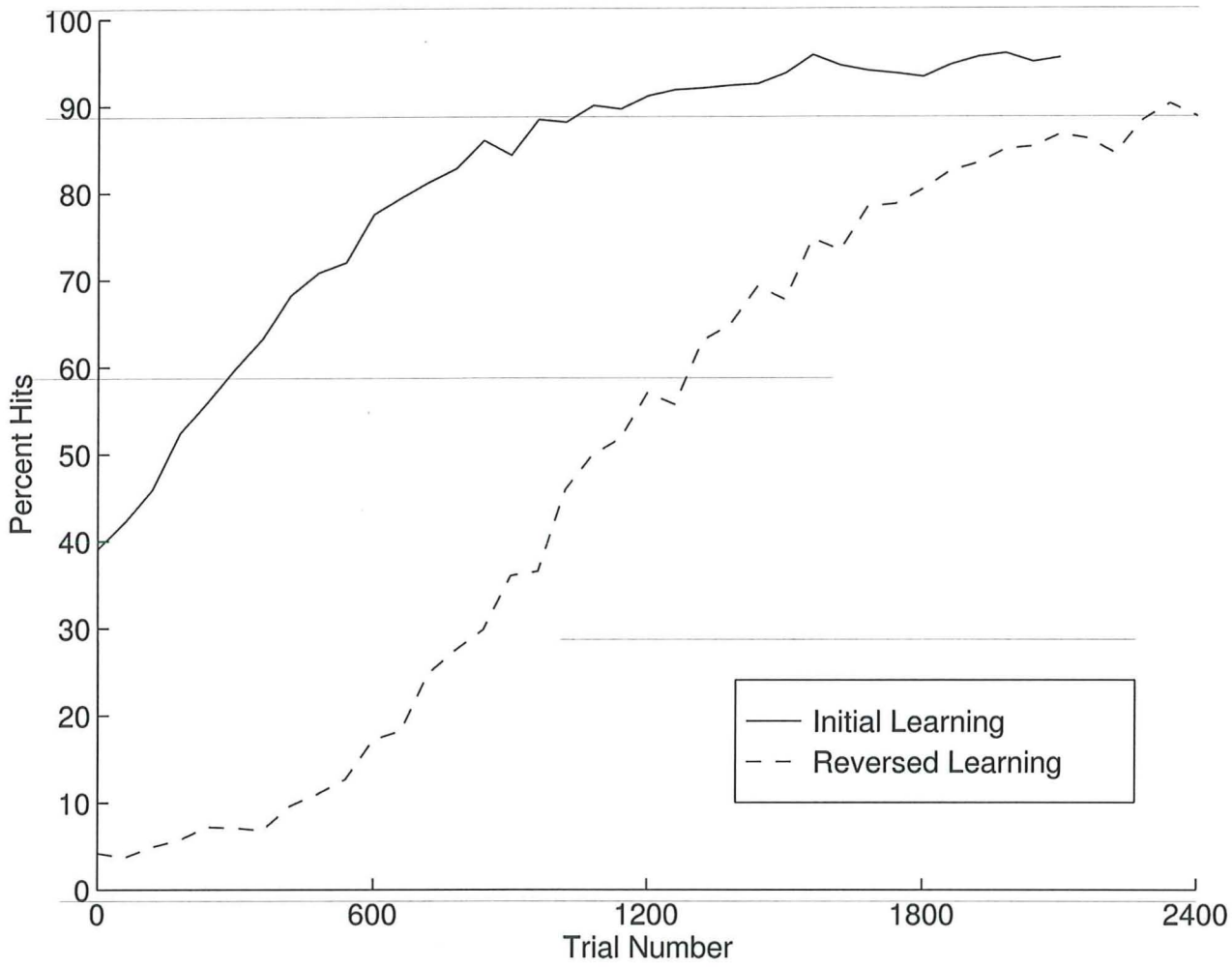


FIGURE 7. Performance of the model in the learning and unlearning paradigm.

reversed and which is strongly dependent on the number of memory items and the frame size. Automatic detection is defined as a parallel process which is difficult to alter, to ignore, or to suppress once learned, and which is virtually unaffected by both the number of memory items and the frame size. In our model controlled search and automatic detection are not two parallel processes, but rather two appearances of one attentional mechanism. The difference in appearances is only due to the different developmental stages of the search system. In the case of automatic detection, the low level visual system is able to indicate the most salient position in the visual field, after much consistent learning. In the case of controlled search this ability has not been learned, and the positions to look at are chosen in some random order. Similar to SLAM (Phaf et al., 1990), our model makes no strict distinction between controlled search and automatic detection, but these are considered to be two extremes on a continuum.

Another explanation of the decrease in reaction time might be that the subjects have learned a new (very complex) feature map. This new feature map may help to direct the focus of attention to the location of the target. Our simulation of the model cannot reject this possibility because the learning of the gating weights given a constant task input is, in a sense, identical to the formation of a higher level feature map. According to Treisman et al. (1992), however, the decrease in reaction time due to learning does not seem to depend on the formation of new feature maps but rather, as in our model, on changes specific for both the task and the stimuli.

One might think that exchanging the targets and distractors would not harm the performance because the learned distinction between these two groups could still be used by changing the classification of target and distractor. However, after the reversal the performance of the subjects drops dramatically and increase only gradually. Like humans, our model has to unlearn the previously learned attention responses and to learn the new relations.

4 GENERAL DISCUSSION

We have proposed a neural network model for selective covert visual attention. This model can learn to focus its attention on important features depending on the task to be fulfilled by modulating the flow of information in the preattentive stage. Computer simulations have demonstrated (1) that the model reveals a performance similar to that of humans in feature and conjunction search after learning, (2) that it can learn visual categories to reduce reaction time without a decrease in performance, and (3) that its learning dynamics are comparable to those of humans.

In all simulations, described in this article, the shared weights of the feature maps ($d^{\theta}_{\mu\nu}$) were fixed. These weights can also be determined through training. However, when using gradient descent as training method complex features appeared to evolve. These features not only were biologically implausible but also made comparisons with psychological data far more difficult. The use of other learning rules than backpropagation can result in biologically plausible feature maps.

Some psychological data can be explained even without any simulations. For example, due to the activation function of the feature map neurons [eqn (2)] our model has a visual search asymmetry like humans (Treisman, 1986). An appealing example of visual search asymmetry is the large difference in reaction time between looking for the presence or absence of a feature. It is easier to find a Q between O's than vice versa. When our model looks for a Q amongst O's the feature map which represents the \ is only active at the location of the Q, resulting in exogenous attention for the target Q. However, when our model searches for an O amongst Q's the feature map representing the \ shows a low (background) activity which is absent at the location of the O. This small difference is not sufficient to drive the focus of attention exogenously towards the target.

During learning no new feature maps are created, but changes are made that are specific for both the stimuli and the task. This explains why the reaction time for learned (automatic) processes substantially increases when the items are slightly changed, e.g., if the contrast of the items is reversed (Treisman et al., 1992).

There are several ways to expand our model in order to explain even more psychological observations. For example, currently, only shape information is used to direct the focus of attention. Attention towards a particular spatial position would require an additional module. This module may either influence the flow of information from the priority map to the saliency map or it can directly affect the saliency map. Similar to humans, "two different attentional systems serve as sources of activation for color or form ... and for location ..., although both might enter and amplify activity within the visual system at the same site" (Posner and Dehaene, 1994).

After learning, our model has a performance similar to that of humans in feature and conjunction search (see Section 3.1). Moreover, the performance is similar to that of the Guided Search model (Wolfe et al., 1989; Wolfe, 1994) and feature-integration theory (Treisman and Gelade, 1980). However, a few differences exist. First, some aspects of visual search mechanisms have not been specified for our model, but are well described in, for example, the Guided Search model. For example, although humans stop searching when they no longer expect to find the target, our model only ends its visual search when all positions have been visited or when the target has been found. Another difference is that our model cannot explain local interactions, such as perceptual grouping (Treisman, 1982), the activation function of the neurons in a feature map [see eqn (2)], which depends on the global total input of this feature map, should be replaced by an activation function like

$$f^{\theta}_{\kappa\lambda} = \frac{i^{\theta}_{\kappa\lambda}}{\mathrm{MAX}\left(1, \sum_{\mu\nu} \left| i^{\theta}_{\kappa+\mu,\lambda+\nu} \right| D_{\mu\nu}\right)} , \tag{8}$$

where $D_{\mu\nu}$ is a monotonically decreasing function of μ and ν (e.g. a Gaussian function). Moreover, in contrast to Guided Search and feature-integration theory, our model reveals similar task-dependencies as humans in the paradigm of Jonides and Gleitman (1972) and showed comparable learning characteristics with those of humans in the experiment of Schneider and Shiffrin (1977).

Although a serial self-terminating process can easily explain that the reaction time is a linear function of the frame size and that the slope of this function on target absent trials is twice the slope on target present trials, a parallel-process interpretation cannot be ruled out (Townsend, 1971). Our model does not depend on a serial process, although in the current simulations the resources of our model "are deployed from item to item in decreasing order of the likelihood that they contain a target. Much the same result could be obtained by deploying a limited resource to all items with the *amount* of resource deployed at each locus dependent on the likelihood that the locus contains a target" (Wolfe, 1994).

In our simulations, we assumed that the information inside the focus of attention is further processed while all other information is rejected, but this division may not be so clear cut. Probably, the farther away from the center of the focus of attention, the more the information is suppressed (or less enhanced). Whether or not this suppressed information reaches the highest cognitive level might depend on whether or not the neuronal responses in this area are synchronized with the responses in the center of the focus of attention. Attention will then be bound to objects because individual objects in a scene are labeled by synchronous neural activity (Engel et al., 1991).

Through computer simulations we have shown that the different performances in several visual search studies (Quinlan and Humphreys, 1987; Treisman and Sato, 1990; Wolfe et al., 1989) can be explained by the difference in usefulness of top-down information. Furthermore, we showed that within the paradigm of Jonides and Gleitman (1972) our model has similar dependence on the interpretation of the task as humans. Finally, the simulations suggest that controlled search and automatic detection need not be two parallel processes (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977) but may also be two different developmental stages of one attentional system.

References

Ahmad, S. (1992). VISIT: A neural model of covert visual attention. In Moody, J. E., Hanson, S. J., and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems 4: Proceedings of the 1991 Conference*, pages 420–427, San Mateo. Morgan Kaufmann.

Amari, S.-I. (1977). Dynamics of pattern formulation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Anstis, S. M. (1974). A chart demonstrating the variations in acuity with retinal positions. *Vision Research*, 14:589–592.

Cave, K. R. and Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, 22:225–271.

Engel, A. K., König, P., and Singer, W. (1991). Direct physiological evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences USA*, 88:9136–9140.

Enns, J. T. (1990). Three-dimensional features that pop out in visual search. In Brogan, D., editor, *Visual Search: Proceedings of the First International Conference on Visual Search*, chapter 4, pages 37–45. Taylor & Francis.

Fukushima, K. and Imagawa, T. (1993). Recognition and segmentation of connected characters with selective attention. *Neural Networks*, 6(1):33–41.

Glasius, R., Komoda, A., and Gielen, S. (1994). Population coding in a neural net for trajectory formation. *Network: Computation in Neural Systems*, 5(4):549–563.

Humphreys, G. W. and Müller, H. J. (1993). SEarch via Recursive Rejection (SERR): A connectionist model of visual search. *Cognitive Psychology*, 25:43–110.

Johnston, W. A. and Dark, V. J. (1986). Selective attention. *Annual Review of Psychology*, 37:43–75.

Jonides, J. and Gleitman, H. (1972). A conceptual category effect in visual search: O as letter or as digit. *Perception & Psychophysics*, 12(6):457–460.

Klein, R. (1988). Inhibitory tagging system facilitates visual search. *Nature*, 334:430–431.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551.

Leow, W. K. and Miikkulainen, R. (1991). A neural network for attentional spotlight. In *Proceedings of International Joint Conference on Neural Networks (Singapore)*, pages 436–441.

Niebur, E. and Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Conference*, pages 802–808, Cambridge. MIT Press.

Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700–4719.

Phaf, R. H., van der Heijden, A. H. C., and Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, 22:273–341.

Posner, M. I. and Dehaene, S. (1994). Attentional networks. *Trends in Neurosciences*, 17(2):75–79.

Postma, E. O. (1994). *SCAN: A Neural Model of Covert Attention.* PhD thesis, Rijksuniversiteit Limburg, Wageningen.

Quinlan, P. T. and Humphreys, G. W. (1987). Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception & Psychophysics*, 41(5):455–472.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological Review*, 84(1):1–66.

Shiffrin, R. M. and Schneider, W. (1977). Controlled and automatic human information processing: II. perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2):127–190.

Townsend, J. T. (1971). A note on the identifiability of parallel and serial processes. *Perception & Psychophysics*, 10(3):161–163.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):194–214.

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 255(5):106–115.

Treisman, A. and Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3):459–478.

Treisman, A., Vieira, A., and Hayes, A. (1992). Automaticity and preattentive processing. *American Journal of Psychology*, 105(2):341–362.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136.

von der Malsburg, Ch. and Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics*, 54:29–40.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238.

Wolfe, J. M., Cave, K. R., and Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433.

Yarbus, A. L. (1967). *Eye Movements and Vision.* Plenum press. trans. B. Haigh.

Nomenclature

$r_{\kappa\lambda}$	the activity of the retina at location (κ, λ)
$i^{\theta}_{\kappa\lambda}$	input of the neuron (κ, λ) of feature map θ
$f^{\theta}_{\kappa\lambda}$	the activity of the neuron (κ, λ) of the feature map θ
$p_{\kappa\lambda}$	the activity of the neuron (κ, λ) of the priority map
$t_{\kappa\lambda}$	the desired output of the neuron (κ, λ) of the priority map
$d^{\theta}_{\kappa\lambda}$	shared weight of feature map θ and relative position (κ, λ)
$a^{\theta}_{\kappa\lambda}$	the attentional shared weight of feature map θ and relative position (κ, λ) (the output of the attentional network)
$a_i(\boldsymbol{m}, \boldsymbol{w})$	the i-th output of the attentional network with input $\boldsymbol{m}$ and weights $\boldsymbol{w}$ [i is short for $(\theta, \kappa, \lambda)$] (the attentional shared weights)
$A_x(A_y)$	the number of receptors in the receptive field of the neurons in the priority map in the horizontal (vertical) direction
$D_x(D_y)$	the number of receptors in the receptive field of the neurons in the feature maps in the horizontal (vertical) direction
$D_{\mu\nu}$	monotonically decreasing function of μ and ν
$\boldsymbol{m}$	the memory input (also called task input)
ξ	Gaussian noise
E	the energy function
q	pattern index
η	the learning parameter
MAX	the maximum function
x_{α}	a feature ($\alpha \in [1, 2, .., 9]$)
y_{β}	a feature either feature x_{β} or feature $x_{\beta+3}$ with $\beta \in [1, 2, 3]$
z_{β}	a feature either feature x_{β} or feature $x_{\beta+6}$ with $\beta \in [1, 2, 3]$
A	a category whose elements have the following features: either 1 or 4 and either 2 or 5 and either 3 or 6
B	a category whose elements have the following features: either 1 or 7 and either 2 or 8 and either 3 or 9
RT	the model's reaction time